Video Surveillance for Road Traffic Monitoring

Abstract

This paper presents a study on road traffic monitoring using the AI City dataset. The goal is to develop a system that can accurately detect and track multiple targets using single-camera and multi-camera techniques. The proposed approach leverages state-of-the-art deep learning models such as Mask R-CNN, YOLOv3 and SSD-512 for object detection and methods like overlap tracking, Kalman filter and Neighbourhood Component Analysis for tracking. Additionally, the proposed system can provide real-time traffic information, which can be used for traffic management and control. Overall, this study demonstrates the potential of AI techniques for road traffic monitoring and provides a useful framework for future research in this area.

1. Motivation

The motivation for this project is centred around the AI City Challenge [1], specifically Track 1: City-Scale Multi-Camera Vehicle Tracking, which aims to improve transportation systems. With the challenge taking place at CVPR 2022, the project team is motivated to develop a solution that addresses the problem proposed in this track. The authors propose two different tasks, with the main task being the Multi-target multi-camera tracking, which is the primary focus of this project. Ultimately, the goal of this project is to contribute to the development of innovative solutions that can enhance transportation systems and improve the quality of life for individuals living in cities.

2. Introduction

Intelligent transport system (ITS) research has recently gained popularity in both academia and business. The CVPR 2020 AI City Challenge is concentrated on 4 different challenges: vehicle counts by the class at multiple intersections, city-scale multi-camera vehicle re-identification, city-scale multi-camera vehicle tracking, and traffic anomaly detection. This is done to speed up research on the creation of smarter transport systems. The city-scale single and multi-camera vehicle tracking is the focus of this work.

3. Related Work

3.1. The Multi-target single-camera tracking

In the multi-target single-camera tracking problem one of the mainly used approaches is the tracking-by-detection, which involves detecting objects in each frame and then linking them together to form tracks. In our experiments, we rely on the baseline locations provided by popular object detectors such as YOLOv3 [11], SSD512 [6], and Mask R-CNN3 [3]. However, there exist newer object detector versions like YOLO V7-8 [15] which improve the object detection autonomous task in mAP by more than 10 points being 25% faster [8].

For the single-camera approach, many different proposals have appeared in recent years. The most important are: The TC [14] is a system that utilizes a fusion of visual and semantic features to cluster and associate data within a single camera view. To improve accuracy, the system also incorporates a histogram-based adaptive appearance model, which learns the long-term history of visual features for each vehicle target. Finally, Tc utilizes re-identification (re-ID) to match objects across different camera views.

The MOANA [13] is a system that employs spatiotemporal data association and utilizes an adaptive appearance model to handle identity switches that can occur due to occlusion or the presence of targets with similar appearances in close proximity.

The DeepSORT [16] is an online approach that utilizes a combination of deep learning features, Kalman-filter-based tracking, and the Hungarian algorithm.

The TrackFormer [7] leverages recent advances in vision transformers to address occluded, missing, or noisy detection.

The TPM [10] algorithm efficiently combines multiple short sub-trajectories into a long trajectory and, using trajectory context. That reduces missing detection.

3.2. The Multi-target multi-camera tracking

Multi-camera multi-object tracking has gained significant attention in recent years, following the development of single-camera multi-object tracking.

There are different viewpoints to tackle this problem.

Some of them utilized graph-based approaches to associate objects across frames and cameras like Graphs [2] and MTMC with tracklet-to-tracklet [4].

In Feature Features for multi-target multi-camera [12], Ristani and Tomasi proposed an adaptive weight loss and hard-identity mining scheme to improve feature learning, while Huang et al. [5] introduced a trajectory-based camera link model that incorporated deep feature re-identification. These methods used the TrackletNet Tracker (TNT) to generate moving trajectories and the camera link model to constrain object order based on spatial and temporal information.

In "A unified multi-view multi-person tracking framework" [17], Yang et al. proposed to join monocular 2D bounding boxes information with 2D poses to produce ro-

bust 3D trajectories to track along multi-camera with overlapping views.
Moreover, Nguyen et al. in "Lmgp: Lifted mul- ticut meets geometry projections for multi-camera multi- object tracking" [9] proposed another approach based on a spatial-temporal lifted multi-cut formulation employing 3D geometry projection.

4. Dataset

In this study, we utilized the AI City Challenge dataset in the CVPR 2020 to investigate traffic behaviour in a mid-sized U.S. city. The dataset consists of 3.58 hours of video footage captured by 46 cameras installed in 16 different intersections. To ensure a comprehensive analysis, we specifically selected sequences 1, 3, and 4 from the dataset. Additionally, it is important to note that the ground truth annotations in the AI City Challenge dataset do not include parked cars.

5. Methodology

In this section, we describe the proposed approach for multi-target single-camera tracking and multi-target multi-camera tracking. We discuss the specific techniques used for each of them and provide details on the pipeline.

5.1. The Multi-target single-camera tracking

In the proposed approach the authors divided the Multi-target single-camera tracking task into three different parts as shown in Figure 1.

5.1.1 Object detection

The first task is object detection, which can be performed with many different approaches like traditional ones: with SIFT SURG, and HoG descriptors. Or with learning techniques like region proposal networks (Mask R-CNN, Faster R-CNN), YOLO... As the presented challenge is assumed to face the AI City Challenge at CVPR 2022 dataset, this task is already solved by the given data. The challenge proposes to use a set of object detection obtained from three different object detectors:

- YOLOv3
- Mask R-CNN
- SSD512

5.1.2 Object tracking

The object tracking step tries to group the object detection from the previous step from the different frames to define the movement of the object along the sequence. In our approach, two main techniques have been studied.

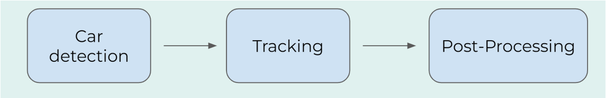


Figure 1. Pipeline for Single-camera tracking

Maximum overlap uses the Intersection over Union (IoU) metric to match bounding boxes detected in consecutive video frames. In each frame, the bounding boxes are assigned to tracks based on their overlap with the closest (highest IoU) bounding boxes in the previous frame. Bounding boxes that do not intersect with any boxes in the next frame are considered to be the end of their tracks, while those that are not assigned to any track are the beginning of a new track. The bounding boxes in the next frame (N+1) are not modified by this technique.

Kalman filter is a tracking algorithm that operates in two steps. The first step is a prediction that estimates the future position of an object, followed by an update step that refines the prediction based on newly observed data. It is an optimal estimation method for tracking the state of a system assuming that the system behaves linearly with Gaussian noise.

5.1.3 Post Processing

Parked Cars removing: In order to address the issue of parked cars in the ground truth annotations, we implemented a refinement process to remove them from the tracking. This was done by setting a threshold for the amount of movement required for a vehicle to be considered in motion. We conducted a grid search to determine the optimal threshold value, testing values between 0 and 1000 with a step size of 25.

To perform the object detection, we used YOLO V3 with camera 010 from sequence 3. Based on the grid search results, we selected a threshold of 550 to remove parked cars from the tracking. This refinement process allowed us to obtain more accurate and reliable results for the object detection and tracking algorithms in our analysis.

Box refinement: To improve the quality of detections, a refinement method has been introduced. This refinement method predicts the future position of a given object by interpolating its tracked detections, matches the object boxes detected in the next frame, and then refines the matched box by computing its centre as the midpoint between the predicted position and the detected one. This refinement method aims to adjust any misaligned detections between frames, provide smoother tracking, and compensate for temporarily lost objects.

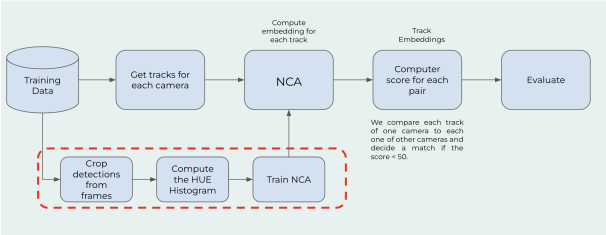


Figure 2. Pipeline for Multi-camera tracking

5.2. The Multi-target Multi-camera tracking

Multi-target multi-camera tracking is a computer vision task that aims to track multiple objects across multiple cameras simultaneously. This task is often encountered in surveillance and security applications, where the objective is to monitor and track the movement of people or vehicles across different camera views. It involves two main tasks object detection and object tracking. Object detection algorithms are used to detect the presence of objects in each camera view. Object tracking algorithms are then used to track the detected objects across different cameras. For object detection, we have used backbones like mask rcnn, single shot detectors and YOLOv3 to detect cars in each camera sequence. For object tracking, we use NCA or neighbourhood component analysis trained on hue histogram features. The basic Pipeline(2) we have used is as follows:

- Use Object detection backbones and get the detections for each frame.

- Crop the detections and extract Hue histogram features and train an NCA model on them.

- Loop over the cameras and get tracks for each camera. For each track, compute the embeddings using the trained model.

- Compare each track of one camera to each one of the other cameras and decide a match if the score < 50 and update the tracks with the assigned detections and evaluate with the required metrics.

6. Evaluation

In this section, we present and discuss the results obtained from the evaluation of the multi-target single-camera tracking using different object detection models. We have evaluated three popular object detectors: Mask R-CNN, SSD 512, and YOLO. The performance of these models has been analyzed for different camera perspectives and sequences using two evaluation metrics, IDF1 and HOTA.

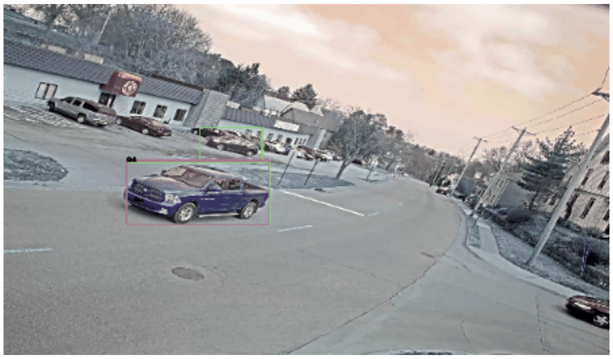


Figure 3. Qualitative results for the multi-target single-camera tracking. Green bounding boxes represent the ground truth.

6.1. The Multi-target single-camera tracking

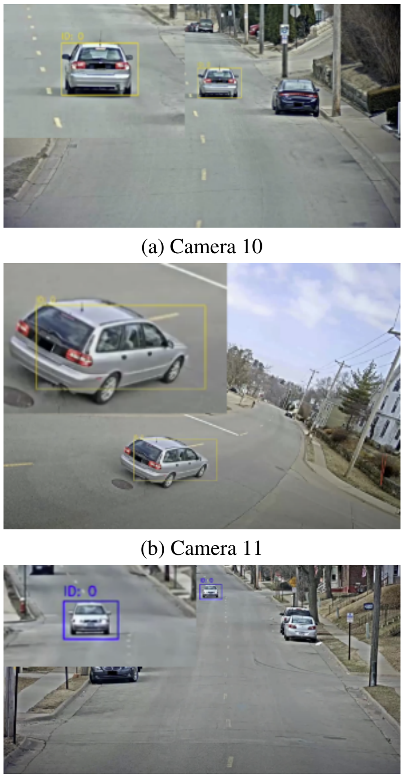
Table 1 shows the IDF1 results for SEQ 3. It is evident from the table that YOLO and Kalman filter-based approaches demonstrate superior performance in terms of the average IDF1 score. The Kalman-YOLO combination achieves the best performance in c11, c13, and c15, whereas the Kalman-SSD 512 combination outperforms the others in c10 and c14. This suggests that the choice of the object detector and the tracking method has a significant impact on the tracking performance, and there is no one-size-fits-all solution.

Similarly, Table 2 presents the results of the HOTA metric for SEQ 3. Here, we observe that the Overlap-YOLO combination delivers the highest average HOTA score. It is interesting to note that although YOLO performs well in the IDF1 metric, the combination with the Kalman filter does not produce the best results in the HOTA metric. This highlights the importance of considering multiple evaluation metrics while selecting an object detection and tracking method for practical applications.

In addition, we also analyzed the qualitative results of the object tracking methods on SEQ 3. We observed that the performance of the tracking methods varies depending on the distance of the cars from the camera and the occlusion of the cars. As shown in Figure 3, we obtained good results when the cars are near the camera and have a clear view. However, when the cars are farther away or occluded by other objects, the tracking methods tend to miss detect the cars, leading to a decrease in the tracking performance.

6.2. The Multi-target Multi-camera tracking

The performance of the object detectors is further analyzed for different sequences, as shown in Tables 3, 4, and 5. The results indicate that the choice of the object detector can significantly influence tracking performance across different sequences. In Seq 1, SSD outperforms the other two detectors in terms of IDP and Precision, whereas



(a) Camera 10

(b) Camera 11

(c) Camera 14

Figure 4. Qualitative results for multi-target multi-camera tracking where the car has been correctly identified as 0 in the three cameras.

YOLO achieves the highest IDR and Recall scores. In Seq 3, YOLO shows greater performance in IDF1, IDR, IDP, and Recall, while SSD achieves the highest Precision score. In Seq 4, YOLO outperforms the other detectors in all metrics except for Precision, where SSD attains a higher score.

These results underline the importance of carefully selecting an object detection model based on the desired performance metrics and the specific characteristics of the application scenario. Moreover, the results suggest that further research is necessary to develop more robust and versatile object detection and tracking methods that can consistently deliver high performance across different camera perspectives, sequences, and evaluation metrics.

Furthermore, we also analyzed the qualitative results of the multi-target multi-camera tracking method, as shown in Figure 4. We observed that, despite the changes in camera perspective, the tracking methods were able to successfully detect the car. However, similar to the single-camera tracking method, we encountered difficulties in detecting cars that were far away from the camera or occluded by other objects. Additionally, we also observed some misclassifications due to the use of hue as a descriptor of a car.

7. Conclusions

In this study, we have investigated the performance of multi-target single-camera and multi-target multi-camera tracking systems using various object detectors, namely Mask R-CNN, SSD, and YOLO. Our experimental results have demonstrated that the choice of object detector plays a significant role in the performance of the tracking systems across different sequences, camera perspectives, and evaluation metrics.

In the case of multi-target single-camera tracking, YOLO exhibited superior performance in several instances, particularly in Seq 3. However, no single approach consistently outperformed the others across all scenarios and metrics. This observation emphasizes the importance of selecting an appropriate object detection model based on the desired performance metrics and the specific requirements of the application.

For multi-target multi-camera tracking, our analysis has shown that the performance of the object detectors can be significantly influenced by the choice of the tracking method. Here too, YOLO demonstrated better performance in some cases, but there was no single method that consistently outperformed the others in all metrics and sequences.

These findings highlight the need for ongoing research in the field of object detection and tracking to develop more robust and versatile solutions for real-world applications. As the performance of the tracking systems can be significantly impacted by the choice of object detector and tracking method, further research should focus on developing adaptive systems that can select the most suitable model and method for the given scenario.

In conclusion, our study has provided valuable insights into the performance of various object detection and tracking methods in the context of multi-target single-camera and multi-target multi-camera tracking systems. We hope that these findings will serve as a basis for future research and development of more advanced and versatile object detection and tracking techniques for a wide range of applications.

References

[1] AI City Challenge. Ai city challenge. https://www.aicitychallenge.org/. Accessed: May 2, 2023. 1

[2] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang. An equalized global graph model-based approach for multi-camera object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11):2367–2381, 2016. 1

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[4] Yuhang He, Xing Wei, Xiaopeng Hong, Weiwei Shi, and Yihong Gong. Multi-target multi-camera tracking by tracklet-

to-target assignment. *IEEE Transactions on Image Processing*, 29:5191–5205, 2020. 1

[5] Hsiang-Wei Huang, Cheng-Yen Yang, Samartha Ramkumar, Chung-I Huang, Jenq-Neng Hwang, Pyong-Kun Kim, Kyoungoh Lee, and Kwangju Kim. Observation centric and central distance recovery for athlete tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 454–460, 2023. 1

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 1

[7] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 1

[8] Upesh Nepal and Hossein Eslamiat. Comparing yolov3, yolov4 and yolov5 for autonomous landing spot detection in faulty uavs. *Sensors*, 22(2):464, 2022. 1

[9] Duy MH Nguyen, Roberto Henschel, Bodo Rosenhahn, Daniel Sonntag, and Paul Swoboda. Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8866–8875, 2022. 2

[10] Jinlong Peng, Tao Wang, Weiyao Lin, Jian Wang, John See, Shilei Wen, and Erui Ding. Tpm: Multiple object tracking with tracklet-plane matching. *Pattern Recognition*, 107:107480, 2020. 1

[11] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[12] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018. 1

[13] Zheng Tang and Jenq-Neng Hwang. Moana: An online learned adaptive appearance model for robust multiple object tracking in 3d. *IEEE Access*, 7:31934–31945, 2019. 1

[14] Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, and Jenq-Neng Hwang. Single-camera and inter-camera vehicle tracking and 3d speed estimation based on fusion of visual and semantic features. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 108–115, 2018. 1

[15] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 1

[16] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1

[17] Fan Yang, Shigeyuki Odashima, Sosuke Yamao, Hiroaki Fujimoto, Shoichi Masui, and Shan Jiang. A unified multi-view multi-person tracking framework. *arXiv preprint arXiv:2302.03820*, 2023. 1

A. Annex 1

In this section, we present the evaluation results of the single-camera and multi-camera object tracking methods. We analyze the performance of different object detection models and tracking algorithms across various evaluation metrics, including IDF1, HOTA, Precision, Recall, IDP and IDR.

	IDF1 (SEQ 3)						Average
Camera	**c10**	**c11**	**c12**	**c13**	**c14**	**c15**	
Overlap,Mask R-CNN	0.609	0.270	0.611	0.472	0.785	0.202	0.491
Overlap,SSD 512	0.895	0.692	0.631	0.872	0.792	0.0	0.647
Overlap,YOLO	0.749	**0.699**	0.672	**0.873**	0.771	0.221	**0.664**
Kalman,Mask R-CNN	0.584	0.371	**0.732**	0.732	0.681	0.200	0.550
Kalman,SSD 512	**0.896**	0.482	0.721	0.723	**0.842**	0.0	0.610
Kalman,YOLO	**0.896**	0.649	0.642	0.831	0.751	**0.241**	**0.664**

Table 1. Multi-target single-camera tracking for idf1 metric

	HOTA (SEQ 3)						Average
Camera	**c10**	**c11**	**c12**	**c13**	**c14**	**c15**	
Overlap,Mask R-CNN	0.552	0.221	0.409	0.336	**0.654**	**1.000**	0.529
Overlap,SSD 512	0.693	0.269	0.457	**0.716**	0.651	0.000	0.464
Overlap,YOLO	0.693	**0.332**	0.455	0.707	0.635	**1.000**	**0.637**
Kalman,Mask R-CNN	0.552	0.242	0.516	0.373	0.471	**1.000**	0.526
Kalman,SSD 512	0.706	0.214	**0.547**	0.515	0.632	0.000	0.436
Kalman,YOLO	**0.729**	0.256	0.494	0.675	0.524	0.941	0.603

Table 2. Multi-target single-camera tracking for hota metric

Sequences	**Detectors**	**IDF1**	**IDR**	**IDP**	**Precision**	**Recall**
Seq 1	Mask RCNN	31.35	28.91	**34.25**	44.72	**52.97**
	SSD	**37.53**	**44.36**	32.53	**69.64**	51.06
	YOLO	37.00	41.00	33.72	62.89	51.73

Table 3. Multi-target multi-camera: sequence 1 detector comparison for various metrics

Sequences	**Detectors**	**IDF1**	**IDR**	**IDP**	**Precision**	**Recall**
Seq 3	Mask RCNN	45.20	38.35	**55.02**	67.85	**97.32**
	SSD	**49.42**	45.05	54.72	77.97	94.72
	YOLO	48.79	**46.61**	51.18	**84.86**	93.16

Table 4. Multi-target multi-camera: sequence 3 detector comparison for various metrics

Sequences	**Detectors**	**IDF1**	**IDR**	**IDP**	**Precision**	**Recall**
Seq 4	Mask RCNN	47.73	44.08	**57.74**	63.73	**90.45**
	SSD	48.66	51.49	49.92	76.98	80.98
	YOLO	**50.74**	**54.27**	52.31	**77.86**	82.49

Table 5. Multi-target multi-camera: sequence 4 detector comparison for various metrics